Deepfake Defences

Mitigating the Harms of Deceptive Deepfakes

Discussion Paper

Published 23 July 2024

Contents

Section

Overview 3

What is a deepfake? 8

Deepfakes that demean, defraud and disinform 11

Responding to deepfakes: Prevent, embed, detect and enforce 17

Where next for tackling deepfakes? 27

Overview

Deepfakes are forms of audio-visual content that have been generated or manipulated using AI, that misrepresent someone or something.

Public concern about deepfakes has increased substantially in recent months. These fears have been stoked by several high-profile incidents, such as the 'deep nude' images of Taylor Swift that went viral on mainstream social media platforms; the fake audio that purportedly depicted London Mayor Sadiq Khan criticising last year's Armistice Day parades; and the deepfake video call that resulted in a finance worker of a major multinational firm transferring £20m to fraudsters.

While these and other well-known cases feature public figures and major institutions, deepfakes are already doing serious harm to ordinary individuals – whether that is by being featured in non-consensual sexual deepfake videos or falling victim to deepfake romance scams and fraudulent adverts. Ofcom's recent poll on deepfakes[1] found that 43% of respondents aged 16+ and 50% of respondents aged 8-15 believed that they had encountered a deepfake at least once in the last six months.

Against this backdrop, we are seeing multiple efforts to stymie the creation and spread of deepfakes, including through the introduction of new laws and offences. This includes an offence that prohibits the sharing of deepfake intimate images. The tech industry, meanwhile, has announced several steps to tackle the spread of harmful deepfakes, with model developers signing up to watermarking schemes and online services moving to tighten their policies on deepfakes and increase the use of detection and labelling tools.

As the new regulator for online safety, Ofcom is committed to doing its part to curtail the circulation of this malicious content.[2] There will be circumstances where services regulated under the Online Safety Act 2023 ('the Act') will need to address the dissemination of some types of deepfake (though not all types). To effectively regulate this type of content, we need to have a clear understanding of the types of deepfakes that can be created, the types of harm they are most likely to be implicated in, their prevalence, and the merits and limitations of different mitigation techniques. This discussion paper details the findings of our research in this area, drawing on insights from expert interviews and our review of the literature.

Key findings

The advent of GenAI tools is changing the deepfakes landscape.

Deepfakes can take many forms, encompassing audio, video, and image content. Often, they involve showing a real person doing or saying something they are not. However, they can also depict entirely fictional characters, and may even be absent of people altogether, instead presenting staged events such as a fake image of a battlefield in a warzone. While some deepfakes consist of wholly novel content, others take the form of existing content that has been manipulated, for example where a real video has been edited to manipulate a person's mouth movements, and where dubbed audio has been added.

[1] Conducted in June 2024 with children aged 8-15 and teenagers and adults aged 16+.

[2] In this paper we focus on deepfakes within the online safety regime, however, Ofcom also has an interest in deepfakes within its other regimes, including Broadcast Standards.

The first deepfakes were created using Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) – two forms of AI that have existed for many years.[3] However, the use of these technologies has been overtaken by the availability of new tools powered by Generative AI (GenAI), which allow users to create wholly new content that is significantly more convincing and life-like. These tools have also made it easier to create deepfakes, such that anybody with modest technical skills can do so.

Not all content created by this technology is harmful, however. GenAI and related tools can be used to augment TV and film outputs in post-production, enhance photos and videos with everyday filters, and create entertaining or satirical material. In healthcare, we have seen synthetic audio be used to replace the voices of victims of diseases like ALS. In the domain of online safety, meanwhile, synthetic data is being deployed to plug gaps in the training datasets required to build vital forms of safety technology. In our desire to address harmful deepfakes, we should be mindful not to undermine the creation of legitimate and innocuous content.

Deepfakes can demean, defraud and disinform.

We identify three main ways that deepfakes can cause harm: by demeaning, defrauding and disinforming.[4]

- **Deepfakes that demean** are created to humiliate or abuse a victim-survivor by falsely depicting them in a certain manner or performing a certain act, for example sexual activity. The impact of demeaning deepfakes, such as those involving intimate image abuse, is often devastating. The intention of the perpetrator may be to abuse or bully victim-survivors, to extort money[5] or to force the survivor or victim (who is often a child) to share further sexual content.

- **Deepfakes that defraud** are used to deceive a target by misrepresenting someone else's identity. They are primarily used to assist fraudulent behaviour such as scam communications and false advertising. The target of defrauding deepfakes is not necessarily important, and they can be channelled at multiple individuals simultaneously.

- **Deepfakes that disinform** are used to sow distrust or spread disinformation across a large group. They are generally intended to be spread widely across the internet, their objectives being to influence or shape societal opinion on key political or societal issues, such as elections, war, religion, or health.

While the prevalence of deepfakes is difficult to measure, based on the evidence available it appears one of the most common forms of deepfakes shared online is nonconsensual intimate content. Leading campaign group My Image, My Choice (MIMC) estimates that there are now over 276,000 videos of this nature circulating on the most popular deepfake sites, with over 4.2 billion total views. The overwhelming majority of this content is thought to feature women, many of whom suffer from anxiety, PTSD and suicidal ideation because of their experiences.

[3] Generative Adversarial Networks are a type of machine learning model that consists of two neural networks (a generator and a discriminator) that 'compete' against one another to generate realistic synthetic content. Variational Autoencoders are a type of machine learning model which learns patterns in data through compressing it and then reconstructing the original data. It can generate new data, including images and speech.

[4] However, we note that these categories can overlap. For example, female journalists are often the victim-survivors of sexualised deepfakes, which not only demean those featured, but contribute to a chilling effect on critical journalism.

[5] Financially motivated sexual extortion or sextortion is a form of blackmail that involves threatening to publish sexual information, photos or videos about someone.

One important driver behind the increase in deepfakes online is the launch of user-friendly apps that serve to make these GenAI models even more accessible, including standalone chatbots and so-called 'nudify' apps. Added to this is the proliferation of websites that are dedicated to hosting deepfake content, and which share manuals that explain how to create such content. Such non-consensual abuse sites and apps are highly likely to allow users to generate and share illegal content. Their user base has dramatically increased – one site reportedly receives 17 million views a month – and are easily accessed through internet search engines. This has enabled use by a wide group of actors, including schoolchildren who have been found to be creating sexual deepfakes of one another.[6] This ecosystem has been described as a 'deepfake economy'.

Tackling deepfakes requires interventions across the technology supply chain.

Addressing deepfakes is likely to require action from all actors in the technology supply chain – from the developers that create GenAI models and related tools, to the platforms that host this technology, through to the user-facing services that act as spaces for deepfake content to be shared and amplified.

In this paper we identify four routes for actors to mitigate deepfakes:

- ***Prevention*** involves efforts to limit the creation of harmful deepfakes. This can include adopting prompt filters to prevent models being instructed to create certain types of content (e.g., nude content); removing harmful content from model training datasets; and blocking outputs before they are presented to users.
- ***Embedding*** involves attaching contextual information to synthetic media through tools including watermarks, provenance metadata and labels. Many organisations, for example, have now signed up to the Coalition for Content Provenance and Authenticity (C2PA) scheme, which has been described as a 'nutritional' label for content.
- ***Detection*** encompasses efforts to distinguish real from fake content, even where no contextual data has been attached to that content. One means of detecting deepfakes is by using machine learning classifiers that have been trained on known deepfake content.
- ***Enforcement*** involves setting clear rules about the types of synthetic content that can be created and shared on online services. It also involves acting against users that breach those rules, for example through suspending or removing user accounts.

All the above interventions show promise in helping to mitigate the creation and spread of harmful deepfakes. Yet it is important to acknowledge their limitations and weak spots. Prevention measures like prompt filters, for example, need to be used carefully to avoid blocking the creation of legitimate content, such as political satire. Some embedding techniques, meanwhile, will not meaningfully address the harms of demeaning deepfakes. They may also be fragile and susceptible to removal by bad actors, and their successful deployment requires extensive coordination between different stakeholders. As such, actors cannot take a piecemeal approach to deploying individual interventions but will need to consider how they can be deployed collectively as part of an integrated mitigation strategy.

Regulated services have a duty to tackle certain types of deepfake.

The Online Safety Act 2023 ('the Act') requires regulated user-to-user services like social media platforms and regulated search services to, among other things, carry out risk assessments to determine the risk of harm to individuals posed by illegal content or content that is harmful to

[6] Sexual content involving children under 18 constitutes child sexual abuse material.

children on their services;[7] and to prevent or minimise the risks that users of regulated services encounter this content.[8] This includes AI-generated deepfake content, where it is in scope of the regime. The Act also requires services to assess the risks of any in-scope GenAI functionalities that they use (e.g. GenAI chatbots) and to take proportionate steps to mitigate those risks.[9]

Ofcom will seek to ensure that regulated services take the necessary steps to protect their users from deepfakes where they are required to do so. If services fail to meet their duties, we will not hesitate to take enforcement action where needed, which may include issuing fines and implementing business disruption measures.

It is Ofcom's responsibility to ensure that services have the tools they need to understand their duties and to execute these effectively. This includes issuing Codes of Practice, which set out the measures that services – both large and small – can take to ensure compliance. We have consulted on measures included in our draft Codes of Practice for illegal harms (IH) and the protection of children (PoC) which would help services tackle illegal and harmful deepfakes. This includes measures that relate to user verification, user reporting, recommender system design and content moderation.

However, our ambition is always to go further. Over the next year, we will examine in more detail the merits and limitations of the measures discussed in this paper, such as the use of hashing and forensic techniques for deepfake detection. As part of this work, we may consider whether one or more measures may be included in a future Code of Practice or guidance. We have also published the results of a separate investigation we are undertaking into red teaming, a type of evaluation for AI models.

While the Act predominantly applies to 'downstream' platforms that interface with users, we urge firms that sit further 'upstream' in technology supply chains to take equivalent action to address deepfakes. This means committing to some basic, first-principle practices, such as evaluating their models, delaying their release if risks have not been sufficiently mitigated (or gating or removing them in the case of model hosts), and taking appropriate action to block, suspend or otherwise sanction users who breach their rules.

In addition, Ofcom will continue to liaise with the Government to identify potential regulatory gaps in relation to deepfakes and GenAI.

[7] The illegal content and children's risk assessment duties are set out sections 9 and 11 for user-to-user services and sections 26 and 28 for search services.

[8] The safety duties about illegal content and the safety duties protecting children are set out in sections 10 and 12 for user-to-user services, and sections 27 and 29 for search services.

9 We are currently consulting on the recommended measures we propose services take that can help them to identify and mitigate risks of harm, including AI risks where relevant.

Discussion papers

Discussion papers contribute to the work of Ofcom by sharing the results of our research and encouraging debate in areas of Ofcom’s remit. Discussion papers are one source that Ofcom may refer to in discharging its statutory functions. However, they do not necessarily represent the concluded position of Ofcom on particular matters.

This paper is not formal guidance for regulated services. It does not recommend or require specific actions. Please see Ofcom’s website for more information about our online safety consultations.

What is a deepfake?

In this section, we define deepfakes, distinguish them from similar types of content like 'cheapfakes', and consider the impact of GenAI in facilitating their creation.

How can we define deepfakes?

The term 'deepfake' first appeared in 2017 as a way of describing videos that featured the faces of female celebrities imposed on the bodies of pornographic actors. Since then, experts and commentators have used the term to describe a wider variety of content types, including images, videos and even audio.

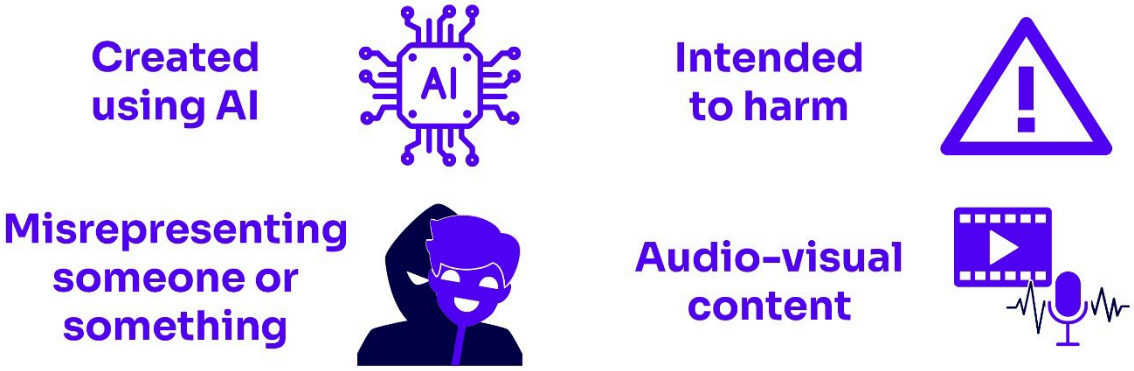
For the purposes of this paper, we define deepfakes as forms of **audio-visual content that have been generated or manipulated using AI, which misrepresent someone or something**.

A key characteristic of deepfakes is that they are usually intended to cause harm by deceiving an audience into believing that something happened when it did not. Recent high-profile examples include the audio forgery of London Mayor, Sadiq Khan, which falsely portrayed him as being disparaging of Armistice Day parades; and the fake video of YouTube influencer MrBeast promoting a scam deal for low-cost iPhones.[10]

Deepfakes, however, do not always feature real people. It was reported that one Chinese state-backed cyber group created deepfake avatars of newsreaders making claims about the fidelity of a pro-sovereignty candidate in Taiwan. Deepfakes may not even feature people at all, but rather an environment or a setting. Deepfakes can be used in conflicts, for instance, to portray bombed out infrastructure and scenes of carnage.

In some cases, deepfakes will constitute wholly new content, whereas in others they take the form of existing content that has been manipulated in some way (e.g. with an object removed or added to a photo).

Figure 1: Defining a deepfake



[10] However, deepfakes can still cause harm without deceiving their audience. For example, an audience may know that a non-consensual intimate deepfake involving a female political candidate is not real, but nonetheless, this deepfake could harm the candidate's reputation, well-being, and sexual autonomy.

We make a distinction between deepfakes and so-called 'cheapfakes', by which we mean deceptive content that is created using simpler methods, like reusing old content for new contexts, or slowing down or accelerating video footage. One of the best-known examples of a cheapfake is a video of Nancy Pelosi that was slowed down at periodic intervals to give the impression she was slurring her words. While cheapfakes are not the focus of this paper, we recognise they can still cause harm and would encourage industry and others to take appropriate steps to address their dissemination.

Synthetic content for good

AI can be used to create benign and innocuous content, just as much as it can be wielded to create harmful content. This includes comedic, self-expressive and satirical material, as exemplified by the image of Pope Francis in a Balenciaga-like jacket, or the numerous synthetic videos featuring Harry Potter characters as viewed through the lens of different cultures.

Synthetic content can also be used for educational purposes. One example is the 'les bleues' video, which used face-swapping technology to highlight misconceptions about women's football.

Outside of these everyday use cases, we have seen synthetic content be used to facilitate industry training, medical treatments and criminal investigations.

At Ofcom, we are considering the potential for synthetic content to aid the development of online safety technologies. A recent report by the Alan Turing Institute explains how AI could be used to create artificial examples of harmful content, in order to train content moderation technology. This is particularly helpful where real versions of that content are rare, for instance shooting incidents that are streamed live.

How has GenAI changed the deepfake landscape?

Bad actors have long used AI tools to create deepfakes. This includes Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).[11] However, the advent of more accessible Generative AI (GenAI) models has materially changed the process of deepfake production. Relative to traditional tools, GenAI is:

- **Simpler and cheaper to use** – It takes a relatively high level of technical expertise to operate GAN and VAE tools, particularly where they are deployed to create video content. In contrast, GenAI tools can be harnessed by anybody with basic technical literacy, with users only needing to enter simple prompts into model interfaces. This in turn lowers the barrier to entry to creating harmful content, allowing new types of actors to participate. Indeed, we have witnessed several cases of children in the UK, US and Spain using GenAI to create deepfake nude images of one another – something that would have been difficult to imagine not long ago.[12] GenAI technology is also cheaper to use. According to one safety technology

[11] Generative Adversarial Networks are a type of machine learning model that consists of two neural networks (a generator and a discriminator) that 'compete' against one another to generate realistic synthetic content. Variational Autoencoders are a type of machine learning model which learns patterns in data through compressing it and then reconstructing the original data. It can generate new data, including images and speech.

[12] Under UK law, this is considered child sexual abuse material.

firm, the cost of cloning a voice has fallen from upwards of $10,000 to just a few dollars in the space of a year.

- **Capable of creating more convincing content** – While deepfakes created using GAN and VAE tools can be high quality, it is much easier to create convincing content using GenAI models (or a combination of the two techniques). A recent study from academics at the University of Waterloo found that only 61 percent of people could distinguish AI-generated images of people and faces from real ones. GenAI tools also continue to improve. The latest video models can be used to create synthetic characters that not only sound lifelike, but which express a range of nuanced facial expressions and body changes, such as subtle head movements.

- **Capable of a greater range of content creation and manipulation** – Whereas GAN and VAE tools are better suited to manipulating existing content, for example in the sense of adding someone else's face into an existing video, GenAI is equally capable of creating entirely new content as well as editing that which already exists. GenAI also allows for more types of content to be created, including long passages of audio, which GAN and VAE tools struggle with. Some of the latest GenAI audio models need only a few seconds of a real person's voice to be able to create an astoundingly life-like clone.

- **Adaptable to specific use cases** – GenAI models are adaptable, especially when they are released on an open-source basis.[13] In such cases, bad actors can 'fine tune' models with additional data, meaning they are more likely to create a specific type of harmful content (e.g. this could involve finetuning an open model to create more sophisticated fraud attacks). Although developers of open-source models may apply safeguards, such as 'nudity' output filters, these can often be easily removed – which in turn increases the ability of bad actors to create harmful deepfakes.

Although some speculate that there is more hype than reality to the claims made about GenAI's capabilities, and that the model development industry could soon be hamstrung by excessive compute costs and a lack of available training data, recent technological advances have undoubtedly transformed the landscape of deepfake production – and in the space of as little as two years.

Figure 2: Summary of the differences between traditional deepfake tools and newer GenAI models

Type of tool	Date available	Capability	Technical skill	Cost
Traditional deepfake tools (GANs, VAEs)	Mid-to-late 2010s	Face swapping, lip syncing, body puppetry, 'nudifying' existing content	Technically complex	Freely available or via paid-for software packages or applications
GenAI models	Early 2020s	Create and manipulate image, video, and audio content	Limited technical skill required	Freely available or via subscription

[13] Sites in the deepfake economy can also be powered by closed source models which have had their safeguards bypassed (i.e., jailbroken).

Deepfakes that demean, defraud and disinform

This section gives an overview of how deepfakes can facilitate harm and sets out a three-part typology of deepfakes, covering those that demean, defraud and disinform. We also look at the prevalence of this type of content and how it has been driven by a 'deepfake economy'.

A typology for deepfakes

From non-consensual sexual imagery, through to fake celebrity product endorsements and made-up recordings of politicians, deepfakes can take myriad forms and harm individuals and society in numerous ways. To help make sense of this confusing landscape of content, we have drawn up a three-part typology that groups deepfakes according to their core purpose:

- **Deepfakes that demean** are created to humiliate or abuse a victim-survivor by falsely depicting them in a certain manner or performing a certain act.
- **Deepfakes that defraud** are used to deceive an individual in a way that results in material gain for the perpetrator.
- **Deepfakes that disinform** are designed to sow distrust or disinformation in an entire – or large subset of – a population.

Deepfakes that demean

Deepfakes that demean are intended to discredit, bully, shame, abuse or humiliate a survivor or victim by falsely depicting them in a particular scenario, for example taking part in sexual activity.[14] Research by McGlynn et al found that intimate image abuse, including artificially generated images, often causes an overwhelming "social rupture" that severely disrupts victim-survivors' lives, shifting their sense of self, their identity and their relationships with their bodies and others.

This content may only require a small audience (or even no audience at all, other than the victim) to be devastating. Demeaning deepfakes often involve women and may disproportionately target black and LGBT+ women. High-profile figures like Taylor Swift, Cathy Newman, Cara Hunter, Alexandria Ocasio-Cortez, and Giorgia Meloni are among those who have been targeted in this way. However, demeaning deepfakes have also been used to victimize people outside the public eye, including ordinary members of the public. In Ofcom's recent survey on deepfakes, 14% of respondents aged 18+ who believed they'd seen at least one deepfake in the last six months said that it was a sexual deepfake. Of these respondents, 64% said it was of a celebrity of public figure, 42% said it was a stranger, and 15% said that it was someone they knew. 88% thought that it depicted someone aged over 18 years old, and 17% thought that it depicted someone aged under 18. 6% said that it depicted themselves.

Sometimes the threat of spreading the deepfake beyond the victim-survivor is enough to accomplish the perpetrator's goal, such as fear or extortion. Moreover, the availability of a demeaning deepfake

[14] Any deepfake featuring a child engaged in sexual activity would be likely to constitute child sexual abuse material (CSAM).

(e.g. on the open web, social media, or someone's hard drive) represents a lifelong emotional and reputational harm to the victim or survivor, even if the content is no longer accessible online.

Deepfakes that defraud

Deepfakes that defraud are primarily used to assist fraudulent behaviour such as scam communications and false advertising, but they can also facilitate other activity including grooming.

The identify of the victims is often not important to the perpetrators who create and share this content. Defrauding deepfakes could be focused on one person or be designed to deceive a mass audience of thousands or even millions. A well-known example of a defrauding deepfake is the fake advert featuring Martin Lewis that was shared on Facebook, in which he appeared to be asking users to sign up for a non-existent Elon Musk investment. Defrauding deepfakes can also be used in romance scams, with fraudsters using GenAI and related tools to create inauthentic profiles with images of non-existent people. More elaborate romance scams have seen fraudsters take part in live deepfake video calls with their victims. Deepfakes can also facilitate grooming, for example, where perpetrators use GenAI tools to create scripts to communicate with and groom a child.

Compared with demeaning deepfakes, the "harm half-life" of a defrauding deepfake may be short, reflecting the opportunistic nature of this form of deception (i.e., either a deepfake does or does not successfully deceive its target). However, fraudulent deepfakes that have been around for days, months or years may still be effective at defrauding their targets.

Deepfakes that disinform

Deepfakes that disinform are primarily aimed at shaping public opinion on political and social issues, such as those relating to elections, healthcare and cultural topics.

Disinforming deepfakes often aim to discredit a particular individual, such as a political candidate ahead of an election, or create controversy surrounding a particular event, such as a military conflict. An example of the former is the audio deepfake of Slovakian politician Michal Simecka, which purportedly recorded him discussing how to rig the upcoming election. The deepfake was shared on social media platforms just hours prior to votes being cast. In other cases, disinforming deepfakes aim to foment disagreement on contentious topics, unrelated to specific individuals or events. Microsoft, for instance, reported that they had observed Chinese-affiliated actors using AI to create deepfakes on politically divisive issues, including gun violence in the United States.

While many disinforming deepfakes are intended to be shared as widely as possible, some are targeted at select groups of individuals or a community bound by certain beliefs. The Global Network on Extremism and Technology (GNET) claim that synthetic audio and image content is being used by extremist networks to spread propaganda and recruit new people to their cause. They discuss how deepfakes have been used to promote the belief that an Islamic apocalypse is imminent, with synthetic images depicting scenes of 'judgement day, battlefields, and even portrayals of catastrophic events'.

While the impact of a disinforming deepfake may diminish following a particular event, such as an election, some deepfakes in this category may be a persistent source of disruption. Deepfakes that amplify conspiracy theories or perennially hateful and misleading claims can have lasting impacts.

Deepfakes and the UK election

Not long after the UK election was announced in May 2024, journalists and expert commentators began to ask whether deepfakes of a political nature might alter the results. During the following six weeks, we saw a number of deepfakes emerge that purportedly showed political candidates making inflammatory and divisive comments. The BBC reported that a network of X accounts were creating and sharing a fake audio clip of Wes Streeting, in which he appeared to insult fellow candidate Diane Abbott, and a deepfake of candidate Luke Akehurst, which falsely showed him making derogatory comments about his constituents. We also saw several incidents where AI-generated images were used to portray false events. An ABC investigation identified a number of Facebook pages that were described as having the 'hallmarks of a Russian influence operation', one of which included fake images of asylum seekers arriving by boat onto the UK's shores. In Ofcom's recent survey on deepfakes, 49% of respondents aged 16+ and 28% of respondents aged 8-15 who believed they'd seen at least one deepfake in the last six months said that it involved a politician or a political event.

Figure 3: Characteristics of the three deepfake categories

	Demean	Defraud	Disinform
Harm example	Non-consensual intimate image abuse	Fraudulent adverts, romance scams	Political, geopolitical, health or other disinformation
Purpose	Discredit, shame, humiliate, abuse	Financial or other harm	Influence opinion
Intended target	Known individual	One or more individuals	A group, e.g., political movement or society as a whole
Intended audience	Few to many	One (e.g., specific target) or many	Many

The three categories are necessarily simplistic. In reality, there will be cases where a deepfake cuts across multiple categories. Women journalists, for example, are often the victims of sexualised deepfakes, which not only demean those featured but may contribute towards a chilling effect on critical journalism. Elsewhere, the Internet Watch Foundation and the National Crime Agency have flagged cases where perpetrators create or threaten to create deepfake intimate images or CSAM to not just demean their victim-survivors, but to defraud them, for example for the purposes of sextortion.

The prevalence of deepfakes

As well as understanding the harm that deepfakes can cause, it is important to try and quantify their prevalence. This can be challenging, not least because definitions of what amounts to a deepfake vary significantly. Moreover, since they are designed to be deceptive, deepfakes may go largely undocumented and therefore uncounted until they are proven to be false. As such, most evidence of deepfake prevalence is anecdotal and reported as and when deepfake-related incidents occur.

While some major online platforms disclose information on the volume of fake accounts, nude content or synthetic content that they have identified, none publish metrics specifically in relation to deepfake content. We therefore have to rely on a combination of public surveys and third-party research to paint a picture of deepfake prevalence. What is more, much of the information in the public domain relates to the volume of deepfake content online, as opposed to user impressions of that content (i.e. its virality). Yet it is imperative to understand both, given that the overall impact of deepfakes is often determined by the number of people who encounter them.

Notwithstanding these measurement challenges, we do have access to at least some valuable evidence on deepfake prevalence. For example:

- Ofcom's recent poll on deepfakes found that 43% of respondents aged 16+ and 50% of respondents aged 8-15 believed that they had encountered a deepfake at least once in the last six months. 10% of those aged 16+ and 13% of those aged 8-15 believed they had encountered deepfakes *over ten times* in this period.
- A joint survey by the Alan Turing Institute and Oxford Internet Institute in 2024 found younger participants aged 18-25 self-reported the highest exposure to non-consensual and political deepfakes.
- A Channel 4 News analysis of the five most visited non-consensual deepfakes websites in 2024 found that almost 4,000 famous individuals were featured, including female actors and musicians.
- A survey undertaken by Internet Matters in 2023 found that 10 percent of children aged 13-16 had either directly experienced or knew of someone who had experienced being featured in fake nude images or videos.
- Analysis undertaken in 2020 by safety technology firm Sensity found that deepfake bots on Telegram had been used to generate more than 100,000 fake nude images, most of which were targeted at women.
- Research by My Image My Choice found 276,149 intimate image abuse deepfake videos on the top deepfake video sites in 2023, with a total of over four billion views, and with more videos uploaded to these sites than in all previous years combined.
- Digital identity company Onfido reported that the number of attempts to use fraudulent deepfakes to circumvent its identity solutions had increased 3000% between 2022 and 2023. They note that a small number of fraudsters are responsible for the majority of deepfake attacks.
- The Advertising Standards Authority's scam ad alert system – which partners with online ad and social media platforms – reported in 2023 that they had seen an increase in the number of paid-for scam ads featuring deepfake footage of celebrities like Elon Musk and Martin Lewis endorsing cryptocurrency and trading apps.
- Research undertaken by Fenimore Harper found that 143 deepfake Rishi Sunak adverts were shown to over 400,000 people on Facebook in the period between December 2023 and January 2024.

- In its 2023 State of Deepfakes report, online security company Home Security Heroes found a 550% increase in the total number of deepfake videos online between 2019 and 2023, and that sexual deepfakes made up 98% of all deepfake videos, 99% of which targeted women.

While it would be inappropriate to draw definitive conclusions from this limited evidence base, these and other studies suggest that one of the most common forms of deepfake shared online today is nonconsensual intimate content. Most of the available evidence also indicates that women are overwhelmingly the targets – among them celebrities, politicians and other public figures, but also ordinary members of the public.

The deepfake economy

What lies behind the apparent increase in the creation and sharing of deepfakes online? As detailed in the last chapter, one significant driver has been developments in Generative AI, which has reduced production costs and allowed for the creation of more life-like and convincing content. Yet these technological breakthroughs are only part of the story. Another driving trend is the emergence of what might be called a 'deepfake economy', with established individuals and groups seeking to make money from aiding this nefarious activity, and established platforms unwittingly facilitating it.

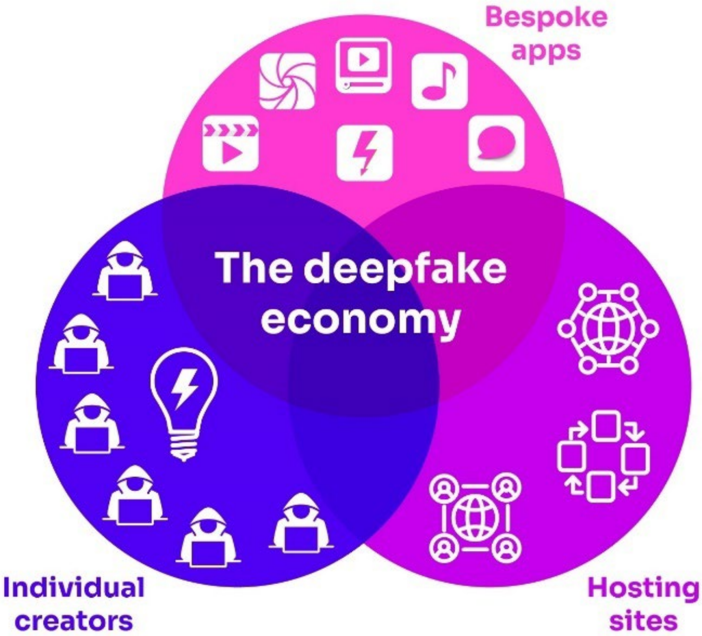
The deepfake economy has several features:

- **Professional creators** – Individuals are now offering to make deepfakes on behalf of others, often for negligible sums. An investigation in 2023 by NBC News found a case of one 'creator' offering to make a 5-minute deepfake for as little as $65. A more recent investigation by 404 Media identified a Telegram user advertising their deepfake services on X for $10. Some of these creators are selling their services on mainstream online platforms, whilst accepting payment through major payment card providers.

- **Bespoke apps** – Several user-friendly apps have emerged that make it easier for people to create deepfakes, with some based on open-source models that are deliberately fine-tuned to create harmful content. This includes over 95 nudify apps, which have different pricing tiers for its users. Another example is a GenAI model which can assist with the creation of fraudulent content like scam emails and messages. Some of these apps are available on mainstream app stores. Indeed, Apple recently removed three nudification apps from its App Store, following pressure from campaigners and journalists.

- **Hosting sites** – Alongside apps that create deepfakes are websites that are dedicated to hosting deepfake content. My Image My Choice believes there are now over 40 such sites in existence. In 2023, one of the most popular deepfake sites was reported as receiving 17 million visitors a month, with some of this traffic coming from internet search engines. As well as hosting deepfakes, these sites often provide forums for deepfake creators to exchange technical know-how and deepfake 'manuals'. These forums have also been used to post job adverts, with creators looking to hire support with creating content.

Figure 4: The deepfake economy



Responding to deepfakes: Prevent, embed, detect and enforce

This section provides an overview and assessment of techniques that can be used by actors across the technology supply chain to address the creation and sharing of deepfakes.

Prevention, Embedding, Detection and Enforcement

What would it take to mitigate the creation and circulation of harmful deepfakes?

Many commentators have pointed to the need for effective detection technology, which can help to distinguish between real and fake content. Yet this is far from the only solution available. We have identified four broad categories of intervention, which can be applied by different actors at different stages in the technology supply chain.

- **Prevention**: This involves efforts to block the creation of harmful deepfakes, with model developers introducing safeguards and adjusting their technology to make it more difficult to create harmful content. Prevention measures include the use of prompt and output filters, and the removal of certain content types from model training datasets.

- **Embedding**: This entails attaching information to content to indicate its origins. This can be done by embedding invisible watermarks on content, attaching provenance metadata to content, and adding labels to AI-generated content as it is being uploaded to platforms. These efforts are primarily undertaken by model developers and online platforms, although they may also involve other actors like cloud computing providers.

- **Detection**: This means using tools to reveal the origins of content, regardless of whether information has been attached to it in the ways described above. These efforts are primarily undertaken by online platforms and can involve the use of both automated and human-led content reviews.

- **Enforcement**: This involves setting and communicating clear rules about the types of synthetic content that can be created using GenAI models and related tools, as well as about the types of content that can be shared on online platforms. It may also involve acting against users that breach those rules, for example by taking down content and suspending or removing accounts.

Figure 5: Four approaches to mitigating deepfakes.

Approach	Purpose	Method	Key actors
Prevention	To limit the creation of harmful deepfakes	Removing harmful content from training data; Using prompt and output filters; Red teaming	Model developers
Embedding	To attach and provide contextual information to content at the point of creation or editing	Content labels; Watermarks; Provenance metadata	Model developers, Online platforms
Detection	To identify a deepfake after it has been created and shared	Forensic techniques; Hashing; User reporting	Online platforms
Enforcement	To set and enforce rules about the types of synthetic content that can be created and shared	Prohibiting content in Terms of Service Account docking; Content downranking	Model developers, Online platforms, Model hosts

Prevention

Prevention measures consist of any attempt to stop a harmful deepfake from being created, and typically involve introducing safeguards 'upstream' to limit what models can produce. Preventative measures include:

Training datasets

Model developers could opt to omit certain types of data from their training datasets. For example, a firm developing an image model could seek to identify and remove "not safe for work" (NSFW) images from training datasets, which could make it more difficult for their technology to generate sexual deepfakes. Similarly, model developers could remove content from their datasets that depicts public figures and celebrities, thereby making it more challenging for users to create deepfakes that portray such individuals.

Prompt filters

Model developers could introduce filters that instruct a model to reject problematic prompt requests. For example, a prompt filter could be set up to reject an instruction to 'create a nude version of this photo'. Midjourney and OpenAI are among the developers that have used prompt filters in this way, including to ban image prompts of political figures like Joe Biden and Donald Trump. The number of terms included in a list of prohibited prompts can vary depending on the specific model and its intended use.

Output filters

In addition to adding filters at the front end of a model where prompts are inserted by the user, model developers can choose to add output filters that automatically inspect generated content and block that which is deemed harmful. For example, a firm developing a text-to-image model could use AI-powered image classifiers to identify and block nude content, which might be used in sexual deepfake imagery. Several cloud computing providers offer to filter the outputs of models that are run using their compute power. This includes Microsoft Azure, which allows customers to use output filters for multiple categories of harm, including sexual and hateful content.

Red teaming

Developers could use red teaming or other methods of evaluation to assess the likelihood of their model creating deepfake content. The results of these exercises can then be used to determine which types of prompts to block, or whether new types of output filter are needed. The findings could also inform a decision on whether to delay the roll out of a model, or to cancel its release altogether. We explore the merits of red teaming in more detail in a separate paper.

Limitations

One attraction of using preventative measures is that it is considerably more efficient to stop deepfake content from being created in the first place than to expend resource in identifying that content once it has begun to circulate online. These measures are particularly well suited to tackling the creation of sexually explicit deepfakes (i.e., deepfakes that demean), as well as some types of defrauding deepfakes where a user's intent to defraud is clear.

However, preventative measures have several limitations:

- **It can be challenging for model developers and other actors upstream to know when a user intends to create content that is harmful.** For example, a user may seek to create an image or video of a celebrity, public figure, or politician for purely satirical purposes. Likewise, a user may want to ask questions of a model that relate to hate and terror for educational reasons. Prompt and output filters can struggle to distinguish between benign requests like these and requests that carry malicious intent. In some cases, it impossible to know whether a given piece of content amounts to a deepfake until it is shared with others, and even then, that judgement call is not straightforward.

- **Preventative measures may be less effective when applied to open-source models**. This is because third-party actors can usually modify these models, including by removing preventative measures installed by the original model developer. In the case of Meta's Llama 2 (an open model), users were able to strip its safeguards to create 'Llama 2 uncensored', which was available for free download on Hugging Face. Third-party actors could also further train (or 'finetune') a model with harmful content, meaning that the model is more likely to create similar content in future. In one widely reported incident, an AI researcher was able to finetune an open source model on content from 4chan's anonymous message board, where users frequently express racist, misogynist and anti-LGBT views. The new version of the model was then used to build several bots, which collectively wrote over 15,000 messages on the same platform in the space of 24 hours.

- **Preventative measures are not always robust**. Even the most well-crafted preventative interventions have weak spots. While filters can stop many attempts to create harmful content, bad actors often still find ways to circumvent them. In the case of the Taylor Swift deepfake incident, while Microsoft Designer had prompt filters to prevent the generation of content featuring public figures, users were able to generate sexually explicit content featuring the singer by misspelling her name in prompts. Similarly, although red teaming can be a valuable way of stress testing models, even the most extensive exercises will not be able to identify every vulnerability in a model, not least because there are a wide variety of 'jailbreaking' techniques that bad actors could deploy.

Embedding

Embedding involves marking content or attaching information about its provenance to indicate whether it is synthetic. Embedding measures include labelling, watermarking, and the application of provenance metadata. Embedding techniques are more likely to be effective against deepfakes that defraud and disinform; proving a deepfake to be false is less effective in addressing the harms of sexualised and demeaning deepfakes.

Labelling

One of the simplest ways of marking content as synthetic is to apply a visible label.

Many online platforms have started to **automatically label content as AI-generated or manipulated**, where they become aware of this. Meta, for instance, recently announced that it would label AI-generated images that users post to Instagram, Threads and Facebook. This builds on the company's existing practice of attaching an 'Imagined with AI' marker to photorealistic images that are created using its Meta AI assistant. Similarly, Snap has announced plans to add a label in the form of a translucent version of the company logo to any content created using its GenAI features, including its My AI chatbot.

Some online platforms are also **allowing or requiring users to self-disclose content as synthetic.** This includes YouTube, which earlier in the year announced a new policy that requires creators to 'share when the content they're uploading is meaningfully altered or synthetically generated and seems realistic'. This information will usually appear in the description fields that sit beneath videos, except in the case of sensitive videos that relate to health, news and elections, where a more prominent label will appear on the content itself.

Watermarking

Watermarking involves adding an imperceptible mark to content that signals whether it is synthetic or genuine. In the case of images and videos, one means of doing this is by making minute alterations to the pixels of the content, in a way that cannot be seen by the naked eye. This is the approach underpinning DeepMind's new SynthID watermarking tool, which the firm claims is resistant to minor content modifications, such as image resizing or colour changes. Once content has been watermarked in this way, it should in theory be detectable using a deep learning algorithm. This detection tool could be deployed either by the model developer or an online platform that wants to detect this content at the point of upload.

In some cases, watermarking methods involve inscribing information onto content that denotes not just whether it is synthetic but who created it, which tools were used to do so, and when it was created. This is known as digital steganography, or the practice of concealing information within larger content files.

While image and video content has been the focal point for watermarking efforts to date, academia and industry groups have begun to develop tools that can inscribe information onto audio content.

One such tool is AudioSeal, which can be used to embed a watermark onto specific sections of an audio file that has been synthetically edited (as opposed to the entire file).

Another area of active study is content watermarking for open-source models. As with providers of proprietary models, many developers that release their tools on an open-source basis (e.g. Stability AI) have added features that automatically embed watermarks onto the content they generate. However, third parties who finetune open-source models can easily strip these functions from the code. To tackle this problem, Meta has proposed an alternative approach called Stable Signature that involves integrating watermarking directly into the image generation process, making it part of the model's core functionality. In theory, this should mean that the model's watermarking instructions remain in place when the model is finetuned. In our conversations with other developers, however, some have suggested this approach can deteriorate image quality.[15]

Metadata

Metadata – or 'data about data' – provides descriptive information about a piece of content, for example about its author, creation or modification date, and the tools used to create or modify it. Unlike watermarking, which involves marking the content itself, metadata is instead added to a file that accompanies the content.

One of the most popular metadata initiatives is the Coalition for Content Provenance and Authenticity (C2PA), which has created a standard for attaching metadata to media content. Under the terms of the standard, metadata is captured and recorded at the point at which content is created, with that information then cryptographically sealed. If the content is subsequently edited (e.g. as images can be within Adobe Photoshop), this creates another layer of metadata which is applied on top of the original. This information can then be viewed using participating software and platforms. Social media platforms could, for example, choose to embed a feature on their feeds that allows users to click on images and videos and see a summary of the corresponding metadata.

A large number of organisations have now signed up to the C2PA standard. This includes camera firms like Canon and Nikon (who can capture metadata at the point where photos are taken), GenAI model developers like OpenAI and Stability AI (who can do the same at the point where GenAI content is created), and online platforms including TikTok and Meta (who can reveal this information to their users, in part by deploying the labels described in the section above).

Limitations

Embedding techniques have the potential to improve our capacity for identifying deepfakes and distinguishing real from fake content. A deepfake video that has been embedded with a watermark at the point of its creation stands a greater chance of being identified as synthetic than one without. Equally, a genuine audio file that has been embedded with provenance metadata is more likely to have its authenticity verified than one that lacks the same inscription.

However, these embedding techniques have their weak spots:

- **Embedding techniques will not be implemented by every model developer or deployer.** As we have seen, some model developers and deployers are deliberately designing tools to create harmful content (e.g. as in the case of nudify apps). It is extremely unlikely that these actors will voluntarily choose to label their content or attach metadata or watermarks.

[15] Moreover, while these watermarking methods are robust to common image transformations such as cropping, their robustness can degrade with more complex manipulations. If an attacker has access to the same auto-encoder used for generation, the watermark can be removed while maintaining high image quality.

- **Bad actors will attempt to remove embedded information.** Those intent on causing harm to others – be it by circulating deepfake adverts or deepfake political content – will always look to eliminate any signal that the content is not genuine. Some metadata fields, for example, can be easily manipulated or removed by bad actors who use file editing or metadata editing tools.

- **Embedding techniques may be less effective for addressing deepfakes that demean**. In the case of sexualised deepfakes or those that contain content intended to bully a victim, supplying contextual information to a viewer may help to prove that the deepfake is false, however the harmful impact of the deepfake can remain the same.

- **Watermarks can be unintentionally weakened through editing.** Content creation and sharing is a messy and convoluted process, often involving multiple rounds of editing, downloading, compression and sharing. Although embedding techniques are becoming more robust, content alterations like these can still damage watermarks and make them harder to detect.

- **Embedding initiatives require close coordination between stakeholders.** A model developer that wants to watermark its content must ensure that other parties (e.g. online platforms) have the tools to detect those same markings. While the C2PA scheme demonstrates that cross-stakeholder collaboration is achievable, these arrangements take considerable effort and time to take root.

- **Popularising embedding techniques could result in genuine content being called into question.** As techniques like labels and metadata become more popular, users of online platforms may expect to see these signals on content as standard and may raise questions where they are not visible. This could lead to perverse outcomes, including cases where authentic content is viewed as fake because it lacks metadata to demonstrate its provenance. Indeed, it may allow individuals who are genuinely depicted in an unflattering circumstance to dishonestly claim that the content in question is a deepfake (a phenomenon known as the 'liar's dividend').

- **More research is needed to understand whether labelling helps users to critically respond to deepfakes.** Consecutive studies have shown that people generally find it challenging to tell fake content apart from that which is entirely or partially synthetic. One study that examined people's ability to spot fake video content found that participants' guesses were, for the most part, 'as good as flipping a coin'. It is possible that media literacy interventions like early years education, content warnings, and community notes on content could improve people's ability to assess more critically what they see and hear online. However, further research is necessary before we can draw robust conclusions. As part of our media literacy duties, Ofcom will be exploring what we can do to support this field of research and practice.[16] We will publish a three-year media literacy strategy in the autumn of 2024, following a period of consultation.

[16] The Online Safety Act (section 165) clarifies Ofcom's existing media literacy duty, which include to help users understand and reduce their exposure to mis- and dis-information.

Detection

Detection refers to techniques that can be used to identify deepfakes, regardless of whether they have been inscribed with markings or attached with information in the ways detailed above. Detection methods are predominantly deployed by online platforms that host content, and can include the use of forensics, hash matching and user reporting.

Forensic techniques

Forensic techniques involve the use of machine learning systems or human review to recognise tell-tale signs that content is wholly or partially synthetic. These techniques vary by content type. For example, to determine whether images are synthetic we could look for a lack of symmetry in facial attributes, as well as erroneous lighting or shadows. For videos, we can look at whether an individual featured in the content is blinking and moving their head in a natural manner. Identifying fake audio content is more challenging but there are still signals that can be monitored, for instance looking for inconsistencies in waveforms that could suggest alterations or tampering.

While humans can perform many of these forensic techniques, they cannot always do so at the speed and scale that online platforms require. To address this challenge, researchers and industry groups have launched several detection tools that promise to automate this process. One such tool is designed to automatically detect changes in colour patterns on a subject's face in a video, which can help to infer blood flow and signal the presence of a real person. Another tool analyses abnormalities in lip movements. Several online platforms have begun to procure and deploy these systems from third parties, including TikTok, which makes use of Resemble AI's technology to monitor audio content.[17]

In addition to these independent tools are those being developed by model developers. OpenAI recently announced that it had developed a classifier to predict whether an image had been created using its Dall-E 3 image generation model. According to OpenAI, the classifier was able to correctly identify 98 percent of Dall-E 3 images (i.e. true positives), while incorrectly flagging only 0.5 percent of non-AI images as coming from Dall-E 3 (i.e. false positives).

Hash matching

Hashing is an umbrella term for techniques that create a 'fingerprint' of a given piece of content. In practice this means using an algorithm to analyse content and create a 'hash' that can represent it. Hashes are then stored in a database that can be accessed by multiple parties as required. In the context of online safety, online platforms can use hashing to notify other platforms of illegal or harmful content they have identified, and vice versa. Hashing databases exist for CSAM,[18] terror content, and non-consensual intimate images. Similar databases could in theory be created for known deepfake content, such as political deepfakes, where that isn't already being captured by existing hashing schemes.

User reporting

As well as relying on their own tools to proactively detect deepfakes, online platforms can invite their users to report this content, which their teams can then review and act on as appropriate. Some online platforms already enable their users to report content which could include illegal or

[17] Other deepfake detection technology companies include iProove, V7 Fake Profile Detector, Deepfake-o-meter, Sensity.

[18] Google has also developed CSAI Match.

harmful deepfake content. Instagram, for instance, enables users to report posts for being 'false information' and for featuring 'scam or fraud' content. Similarly, YouTube allows users to report content that is 'misleading or deceptive with serious risk of egregious harm'.

Limitations

Detection methods like content classifiers and user reporting tools are vital in efforts to tackle deepfakes. This is especially the case where deepfake content lacks watermarks and metadata, as is often the case with content generated by open-source models. Many online platforms are already investing huge amounts in detection interventions, and academic research in this area continues to progress at pace.

However, the successful roll out of detection methods faces several hurdles:

- **Bad actors will always seek to adjust their methods to outmanoeuvre forensic techniques.** This means that researchers and industry groups will need to continuously update their tools and identify increasingly subtle signals that content may be synthetic. One of the forensic experts we spoke with said that forensic techniques tend to have a shelf-life of between 2-5 years before they need to be replaced.
- **Identifying deepfakes requires more than just knowing whether content is synthetic.** It also involves a determination of whether that content is intended to misrepresent. This type of assessment can often only be made by humans, making deepfake detection a costly endeavour. This risks putting some detection methods out of reach of smaller platforms (and wider civil society groups), and could mean that efforts to detect deepfakes featuring ordinary individuals are deprioritised in favour of those targeted at high-profile public figures.
- **Content editing can diminish the accuracy of deepfake detection tools.** Research undertaken by human rights advocacy group Witness shows that making modest adjustments to content can deteriorate the accuracy of detection tools. Testing a variety of tools, they found that lowering the resolution of content, cropping content, and creating duplicates of content often led to incorrect detection results. They also noted that it can be challenging to interpret the results of some tools, particularly where they give binary 'yes/no' results.
- **Hashing techniques can be vulnerable to 'collision'.** This describes circumstances where different content has the same hash value, which could result in genuine content being identified as a deepfake (or vice versa). This phenomenon is particularly problematic where content that has already been hashed is altered or modified to create a deepfake.[19]
- **Users of online platforms can find it challenging to identify deepfake content.** In 2022, the Royal Society found that most people struggle to detect high quality deepfakes, regardless of prior awareness, self-confidence in deepfake detection or internet and social media proficiency. In Ofcom's recent survey on deepfakes, only 9% of respondents aged 16+ said that they were confident in their ability to identify a deepfake if they saw one. In addition, online users often note that reporting mechanisms on online platforms can be confusing, time-consuming and 'unduly laborious', which may discourage them from entering the

[19] In addition, collisions can be random or adversarial. Random collisions are inevitable due to the limitations of compressing large data into smaller hashes, with increasing likelihood as more files are hashed or as length decreases. Adversarial collisions are deliberately caused by slight modifications to the content (i.e., by creating a deepfake) to alter its hash without changing its appearance, which poses a significant challenge to hashing.

process. This is supported by Ofcom research, which found that while 60% of online users encounter harmful content, only 20% of them would report it.

Enforcement

Enforcement involves setting and communicating clear rules about the types of synthetic content that can be created using GenAI models and related tools, as well as about the types of content that can be shared on online platforms. It is often the foundation on which platforms build their safety measures, including their approach to moderating content and usage. It also involves acting against users that breach those rules, for example through taking down content and suspending or removing accounts.

Setting rules

Many online platforms have introduced specific rules within their terms of service and community guidelines that clarify the types of synthetic content allowed on their sites. TikTok, for instance, requires that all 'synthetic or manipulated media that shows realistic scenes' must be clearly disclosed. They also prohibit the posting of synthetic media that contains the likeness of any real private figure', among other content types. X, meanwhile, say they do not allow users to share synthetic, manipulated, or out-of-context media that may deceive or confuse people.

Our review of platform policies shows that most now include specific reference to synthetic or manipulated media, although are often agnostic as to how that content has been created. Moreover, almost all platforms hold deliberate deception as an essential component of prohibited content.

Even in those cases where online platforms do not have policies that relate to synthetic content, this type of media could still be captured by other, more general policies. YouTube, for example, has an impersonation policy that bars the sharing of videos that impersonate someone else, while their nudity and sexual content policy prohibits the sharing of explicit content that is intended to be sexually gratifying. The latter includes content that depicts someone in a sexualised manner without their consent.

Other actors in the technology supply chain have similar policies. Model developer Anthropic has a Usage Policy that forbids several practices, including the use of its models to create content that could aid misinformation, facilitate fraud and incite violence or hateful behaviour. Similarly, model host Hugging Face prohibits the sharing of code and other 'model artefacts' on its platform that could be used to create content that harms others.

Enforcing rules

Each of these actors can take enforcement action where their rules are breached. This includes:

- **Issuing warnings to users** – Users can be issued a warning or a 'strike', notifying them that they have breached the rules. Some platforms offer policy training to their users after an initial warning.
- **Taking down content** – For example, where the content clearly violates a platforms' terms of service.
- **Suspending or removing users** – User accounts can be suspended or docked. For example, OpenAI removed a developer account that created a bot impersonating US presidential candidate Dean Phillips. Some model developers choose to move offending users onto a

restricted version of their service that has more limited capabilities. User accounts can also be terminated entirely.

- **Labelling content where there isn't a clear breach -** Where there isn't a clear breach, online platforms may take a decision to label content or models. In the case of Hugging Face, for content that is not fully prohibited, the platform can request model owners to add a 'Not for All Audiences' tag to their models, or to 'gate' the model to make it less visible to others.

To date, limited information has been disclosed by online platforms, model developers and model hosts about how they have enforced their policies in relation to synthetic content and deepfakes.

Limitations

Establishing and enforcing clear rules makes it less likely that users will be able to create and share deepfake content. Clear terms of service, community guidelines and licence agreements can reduce the ability of bad actors to exploit loopholes, whilst also enabling content moderators to make fairer and more informed decisions as they review content. However, this can be hard to get right:

- **Policies can suffer from arbitrary boundaries.** The rules set by some online platforms appear to be incoherent. Meta's Oversight Board, for example, recently criticised Meta's synthetic media policy as being 'inappropriately focused on how the content has been created', only applying to video content and content that makes people appear to say words they did not say. Meta subsequently expanded their policy in line with the Board's recommendations.
- **Policies can lack specificity.** The terms of service and community guidelines of some firms in the technology supply chain can be too generic, for example prohibiting the creation or sharing of 'harmful content' without providing detailed examples of what that means in practice.
- **Licence agreements are difficult to enforce in the case of open-source models.** Once an open-source model has been released it is difficult to monitor who is using it and for what purposes. Even where a model developer is aware of their models being misused to create prohibited deepfake content, they are rarely able to step in and block that behaviour.

Tackling deepfakes requires a multi-pronged approach.

The interventions detailed in this chapter show promise. But an individual measure taken on its own is unlikely to significantly change the dial in aiding platforms to effectively tackle deepfakes. Actors seeking to address the risks posed by deepfakes on their models or their platforms will likely find that they need to stand up a deepfake mitigation strategy that implements a combination of these interventions.

Where next for tackling deepfakes?

In this paper we have looked at the increase in prevalence and potential for harm from deepfake content online. We have shown that deepfakes can take many forms, and that they can harm adults, children and society in any number of ways – from scamming victims out of money, to disrupting elections, to demeaning people in sexual imagery. As we see further advances in GenAI, the sophistication and scale of this content is only likely to grow.

It is imperative therefore that all players in the technology supply chain take action today to curtail the creation and sharing of this malicious content. This includes services regulated by the Act, as well as non-regulated services that lie outside of the regime.

Our expectations of regulated services under the Act

Ofcom will ensure that regulated services – be they social media platforms, video sharing sites or search engines – take the necessary steps to protect their users from deepfakes where they are required to do so.

It is Ofcom's responsibility to ensure that services have the tools they need to understand their duties and to execute these effectively. This includes publishing guidance on how services can conduct their risk assessments, and issuing Codes of Practice, which set out the measures that services – both large and small – can take to ensure compliance. We have consulted on measures included in our draft Codes of Practice for illegal harms (IH) and the protection of children (PoC) which would help services to tackle illegal and harmful deepfakes. Such measures include: [20]

- **User verification and labelling schemes** – Having clear internal policies for operating notable user or paid-for verification schemes (where those exist) and improving transparency for users about what verified status means in practice. This may better equip users to assess whether a user posting content is authentic or impersonating a high-profile individual or organisation and inform how they respond to it (IH).
- **Recommender systems** – Collecting additional metrics within recommender system tests to understand whether design changes could increase user exposure to illegal content, which could include exposure to some types of deepfake content (IH). We also propose services should design their recommender systems so that they filter out PPC, and downrank or limit the visibility of other forms of content that is harmful to children in their recommender feed (PoC).
- **Content moderation** – Setting performance targets for content moderation functions, adequately resourcing content moderation teams, and training staff on how to identify and take down illegal content, which would likewise include deepfake content (both IH and PoC).

[20] Note that these measures are segmented. Some apply only to the largest and riskiest of services, and other apply to all services.

- **User reporting and complaints** – Establishing complaints systems that are transparent and easy to access and use (both IH and PoC) and establishing a dedicated reporting channel for fraud (IH only), which would enable trusted flaggers to report illegal deepfake fraud content.
- **Search service design** – Enabling users to easily report predictive search suggestions that could direct users to illegal content or content that is harmful to children (both IH and PoC). This could include search suggestions that direct users to sites that host deepfake content.

In due course, we will publish our IH and PoC statements confirming the final measures in our Illegal Harms and Children's Safety Codes of Practice and our final guidance.

Yet there is no excuse for these services to sit idle in the meantime. We encourage them to act now – as some are already doing – to protect their users from encountering and being targeted by this content. If services fail to meet their duties, we will not hesitate to take enforcement action where needed, which may include issuing fines and implementing business disruption measures.

What we will do next on deepfakes and GenAI

These measures provide a strong starting point, but our ambition is always to go further. Over the next year, we will further explore and assess the merits and limitations of the measures discussed in the previous chapter and consider them in our future policy work. In particular, we will:

- Examine the role of deepfakes in facilitating fraud offences, which will inform our forthcoming Fraudulent Advertising Code.
- Explore measures that can address synthetic CSAM, some of which could fall under our definition of a deepfake.
- Examine the role of deepfakes in facilitating online gender-based violence and abuse, which will inform our forthcoming guidance for services on protecting women and girls.
- Publish the findings of our research into red teaming, a type of AI model evaluation which can help to identify safety vulnerabilities.
- Continue to engage with UK users to monitor how they experience deepfakes online and their wider use and attitudes towards GenAI, as well as continue to support media literacy research and initiatives.

In addition, Ofcom will continue to liaise with the Government to identify potential regulatory gaps in relation to deepfakes and generative AI.

What non-regulated services can do

The Act predominantly applies to 'downstream' services that interface with users, rather than 'upstream' actors like model developers and model hosts (with some exceptions). However, as we have seen, it will be challenging to make inroads into tackling deepfakes without intervening upstream at the point at which models are designed, finetuned and accessed. We encourage model developers and model hosts to implement appropriate safeguards to prevent their technology from being misused, and to work with online platforms to help them identify deepfakes produced or facilitated using their products.

While there are limitations to some of the measures introduced in the last chapter, we suggest there are several first principle steps that model developers and model hosts can follow to better safeguard the end-users of their models and those who may encounter content created by them:

- Conducting red teaming exercises (or other evaluation methods) on the models they build or host, to assess the likelihood of them creating deepfake content.

- Delaying the release of models where these evaluations reveal safety risks, and potentially halting the release of models entirely where these risks cannot be sufficiently mitigated.
- Making use of prompt and output filters, where these are proven to be effective and do not disproportionately limit the creation of benign content.
- Ensuring they have conducted appropriate checks on the data used to train models, removing any problematic content that might aid the creation of deepfakes.
- Investigating the value of adding provenance metadata and watermark functionalities to models, noting the caveats set out in the last chapter.
- Communicating to users how models can be used (in the case of developers) and what models can be shared with others (in the case of model hosting sites).
- Like online platforms, take appropriate action to block, suspend or otherwise sanction users who breach these rules, being mindful to uphold user rights to freedom of expression.

We will continue to engage widely with experts on deepfakes, and we welcome your feedback on the findings and arguments in this paper. Contact our Technology Policy team at TechnologyPolicy@Ofcom.org.uk.